ECGECHODIAG: A BENCHMARK FOR ECG-CONDITIONED MULTI-VIEW ECHOCARDIOGRAPHY GENERATION AND CARDIAC DISEASE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiography (ECG) enables prolonged cardiac monitoring but cannot directly visualize cardiac anatomy or motion. ECG-conditioned echocardiography (Echo) generation offers a potential way to infer this information, yet whether generated videos preserve complementary diagnostic evidence across views remains unclear. We introduce ECGEchoDiag, a dataset and benchmark for ECG-conditioned multi-view Echo video generation and evaluation of diagnostic information preservation. It contains 15,715 synchronized ECG–Echo cardiac cycle pairs from 991 patients across 10 views, including 507 patients with multi-label annotations for eight cardiac disease categories. The dataset preserves within-patient associations across views and temporal synchronization within each recording. We further propose EchoGenDiag, which combines target-aware cross-view conditioning, cardiac-phase-dependent motion modeling, and reference-guided video reconstruction. For each target view, the model integrates its reference frame and ECG with information from other available views of the same examination. A frozen diagnostic teacher provides multi-label supervision and prediction consistency constraints to encourage disease-related information preservation. We evaluate reconstruction quality, temporal consistency, and patient-level multi-label disease recognition using patient-disjoint splits. Comparisons between real and generated videos assess diagnostic information preservation, while single-view and multi-view comparisons examine the contribution of cross-view conditioning.

1 INTRODUCTION

Cardiovascular disease imposes a major global health burden Vos et al. (2020), making timely detection of cardiac abnormalities important for assessment and intervention. Electrocardiography (ECG) and echocardiography (Echo) are widely used, complementary modalities: ECG records cardiac electrical activity to assess rhythm and conduction, while Echo directly visualizes cardiac anatomy, valve motion, and myocardial function. Wearable ECG devices facilitate prolonged monitoring outside clinical settings, helping detect arrhythmias that brief examinations may miss Steinberg et al. (2017). However, ECG cannot directly visualize cardiac structure or mechanical motion, while Echo typically requires specialized equipment and trained operators, limiting continuous or frequent acquisition. Consequently, when cardiac conditions change between clinical visits, out-of-hospital monitoring often lacks imaging evidence of these structural and functional changes.

Cardiac electrical activity and mechanical motion are closely linked through excitation–contraction coupling, providing a physiological basis for inferring cardiac mechanics from ECG Bers (2002). Previous studies have used echocardiographically defined abnormalities as supervision to detect left ventricular systolic dysfunction Attia et al. (2019) and multiple structural heart diseases from ECG Ulloa-Cerna et al. (2022); Poterucha et al. (2025). ECHOPulse further extends ECG-based inference from disease labels to video generation, using ECG and reference Echo frames to generate A4C or A2C videos. Its evaluation examines video quality, cardiac phase alignment, and left ventricular ejection fraction (LVEF) consistency Li et al. (2025). However, these assessments do not establish whether generated videos preserve the structural and motion information needed for patient-level multi-label disease recognition.

One promising direction for preserving richer diagnostic information is to exploit the complementarity of echocardiographic views. Clinical diagnosis typically integrates anatomical and motion evidence across views. Recent studies show that multi-view aggregation supports diverse cardiac assessment tasks Holste et al. (2025) and improves abnormality detection over single-view models Barrios et al. (2026). Whether this complementary information is preserved in ECG-conditioned videos and remains useful for disease recognition is still unclear. Addressing this question requires models to reconstruct view-specific anatomy and dynamics while preserving disease features across views and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

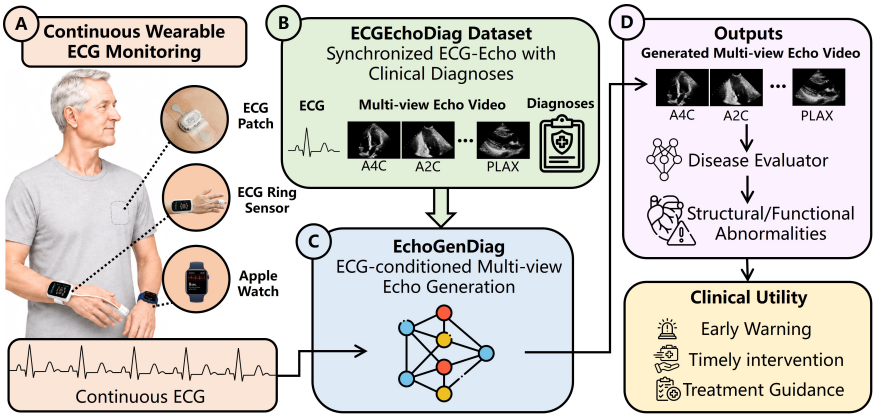


Figure 1: Overview of ECGEchoDiag and EchoGenDiag. (A) Wearable devices enable continuous ECG monitoring. (B) ECGEchoDiag provides synchronized ECG–Echo recordings across multiple views with patient-level disease annotations. (C) EchoGenDiag generates ECG-conditioned multi-view Echo videos. (D) A disease evaluator jointly analyzes the generated views for patient-level disease recognition, with potential applications in early warning, timely intervention, and treatment guidance.

aggregating diagnostic evidence from varying combinations of available views. We therefore ask: *Can multi-view Echo videos generated from ECG and reference frames jointly preserve patient-specific diagnostic information and support patient-level multi-label disease recognition?*

To investigate this question, we introduce ECGEchoDiag (Table 1), a dataset and benchmark for ECG-conditioned multi-view Echo video generation and evaluation of diagnostic information preservation. ECGEchoDiag contains 15,715 synchronized ECG–Echo cycle pairs from 991 patients across 10 echocardiographic views. A diagnostic subset of 507 patients provides patient-level multi-label annotations for eight cardiac disease categories. The dataset preserves within-patient associations across views and ECG–Echo temporal synchronization within each record, while the diagnostic subset supplies disease supervision. Together, these support video reconstruction, cross-view learning, and patient-level disease recognition.

Building on this dataset, we propose EchoGenDiag, a framework for ECG-conditioned multi-view Echo video generation with diagnostic-preservation supervision. Figure 1 summarizes the dataset and framework. For each target view, the model combines its reference frame and synchronized ECG with complementary information from other available views of the same examination through target-aware cross-view fusion. The resulting representations modulate ECG-conditioned dynamics across cardiac phases, while multi-scale reference-feature warping and content completion reconstruct anatomy and motion. To encourage diagnostic information preservation, a frozen diagnostic teacher provides multi-label supervision and enforces consistency between predictions from real and generated videos, using label agreement to select reliable teacher predictions. We evaluate reconstruction quality, temporal consistency, and patient-level multi-label disease recognition using patient-disjoint data splits. Comparisons between real and generated videos assess diagnostic information preservation, while comparisons between single-view and multi-view variants examine the contribution of cross-view conditioning.

Our main contributions are:

- We introduce ECGEchoDiag, a dataset and benchmark providing synchronized ECG–Echo cardiac cycles, within-patient associations across views, and patient-level disease annotations to support multi-view video generation and evaluation of diagnostic information preservation.
- We propose EchoGenDiag, which combines target-aware cross-view conditioning, cardiac-phase-dependent motion modeling, and reference-guided video reconstruction. Diagnostic-teacher supervision encourages generated videos to retain disease-related information beyond visual reconstruction quality.
- We establish an evaluation protocol covering reconstruction quality, temporal consistency, and patient-level multi-label disease recognition. Comparisons between real and generated videos assess diagnostic information preservation, while single-view and multi-view comparisons evaluate the contribution of cross-view information.

Table 1: **Comparison of public ECG and echocardiography datasets.** Record counts use each dataset's native data object. For EchoXFlow, the record and view counts describe its triplane subset. TMED-2 counts "Other" as one of its five view classes.

Dataset	#Patients	#Records	ECG	Echo	Echo data	#Views	Sync.	#Disease classes
PTB-XL (Wagner et al., 2020)	18,885	21,837	✓	✗	–	–	✗	5
MIMIC-IV-ECG (Gow et al., 2023)	161,352	800,035	✓	✗	–	–	✗	Uncurated
EchoNext (Elias & Finer, 2025)	36,286	100,000	✓	✗	–	–	✗	11
CAMUS (Leclerc et al., 2019)	500	1,000	✗	✓	B-mode	2	✗	–
EchoNet-Dynamic (Ouyang et al., 2020)	10,030	10,030	✗	✓	B-mode	1	✗	–
TMED-2 (Huang et al., 2021)	577	17,270	✗	✓	B-mode	5	✗	3
CAMEO (Lee et al., 2025)	100	900	✗	✓	B-mode	9	✗	2
MIMIC-IV-ECHO (Gow et al., 2026)	4,572	524,137	✗	✓	Doppler, B- and M-mode	Uncurated	✗	Uncurated
HeartCycle (Fernandez et al., 2025)	17	208	✓	✓	Doppler, M-mode	–	✓	–
EchoXFlow (Stenhede et al., 2026)	622	573	✓	✓	Doppler, B- and M-mode	3	✓	–
ECGEchoDiag (Ours)	**991 (507 labeled)**	**15,715 (8,494 labeled)**	✓	✓	**B-mode**	**10**	✓	**8**

2 Related Work

2.1 Cardiac Disease Recognition from ECG and Echocardiography

AI-ECG studies use echocardiography-derived labels to predict left ventricular systolic dysfunction (Attia et al., 2019), left-sided valvular disease (Elias et al., 2022), and multiple structural heart diseases (Ulloa-Cerna et al., 2022; Poterucha et al., 2025). These findings suggest that ECG contains diagnostic information associated with structural and functional abnormalities observed on echocardiography. However, these methods primarily output labels or risk scores without explicitly visualizing cardiac structure and motion.

AI-Echo research has expanded from view-specific functional assessment and disease recognition to multi-view diagnosis. Single-view methods estimate LVEF from A4C videos (Ouyang et al., 2020), detect cardiac amyloidosis (Goto et al., 2021), distinguish it from constrictive pericarditis (Chao et al., 2024), and identify regional wall motion abnormalities from parasternal short-axis images (Kusunose et al., 2020). Multi-view approaches address congenital heart disease recognition (Wang et al., 2021) and aortic stenosis grading (Huang et al., 2021; Mokhtari et al., 2023), extending to multiple valvular regurgitation assessments (Long et al., 2025) and broader disease recognition (Holste et al., 2025; Vukadinovic et al., 2026). Direct comparisons further demonstrate diagnostic gains over single-view models (Barrios et al., 2026). These approaches analyze real echocardiograms; we investigate whether ECG-conditioned multi-view videos preserve complementary diagnostic information for patient-level multi-label disease recognition.

2.2 Conditional Echo Video Generation

Conditional Echo video generation controls cardiac appearance and motion using LVEF (Reynaud et al., 2022), reference images and LVEF (Reynaud et al., 2023; 2024; 2025; Oladokun et al., 2026), multimodal cues such as text, segmentation masks, and optical flow (Zhou et al., 2024), curves describing changes in cardiac structure position and size (Yu et al., 2024), and reference frames from arbitrary cardiac phases (Zhang et al., 2026). Disease labels also guide video synthesis for rheumatic heart disease (Roshanitabrizi et al., 2025), atrial septal defects, and pulmonary arterial hypertension (Muhammad et al., 2025) to augment disease recognition training data. However, these methods mainly rely on predefined conditions or existing images, lacking continuous physiological signals reflecting the patient's current state. ECHOPulse generates single-view A4C or A2C videos from patient ECG and a reference Echo frame (Li et al., 2025), demonstrating the feasibility of ECG-driven cardiac motion reconstruction. Since wearable devices enable continuous ECG acquisition, this approach offers potential for inferring cardiac mechanics and supporting long-term monitoring. However, ECHOPulse neither investigates joint multi-view generation nor evaluates whether generated videos preserve complementary disease information across views. We further explore ECG-conditioned multi-view video generation and its value for joint patient-level diagnosis.

2.3 Public ECG and Echo Datasets and Benchmarks

Table 1 compares public datasets in terms of modality availability, view coverage, synchronization, and disease annotations. Public ECG datasets primarily support cardiac electrical abnormality detection, including PTB-XL (Wagner et al., 2020) and MIMIC-IV-ECG, which links ECGs to clinical records (Gow et al., 2023). EchoNext further provides

Echo-derived cardiac disease labels (Elias & Finer, 2025), but lacks corresponding Echo videos and thus cannot directly support paired video reconstruction. Public Echo resources focus on structural, functional, and disease analysis. CAMUS provides A2C/A4C sequences with cardiac chamber annotations (Leclerc et al., 2019), EchoNet-Dynamic offers A4C videos with cardiac function labels (Ouyang et al., 2020), and TMED provides multi-view images with aortic stenosis labels (Huang et al., 2021). MIMIC-IV-ECHO contains multi-view examinations and structured measurements (Gow et al., 2026), CAMEO supports view classification (Lee et al., 2025), and ETAB provides evaluation protocols for segmentation, classification, and functional estimation (M Alaa et al., 2022). Synchronized resources connect electrical signals with mechanical activity. HeartCycle provides synchronized ECG and M-mode/Doppler Echo from healthy participants for hemodynamic research (Fernandez et al., 2025). EchoXFlow provides synchronized ECG and multimodal Echo for blood flow and segmentation studies but lacks diagnostic labels (Stenhede et al., 2026). Existing benchmarks do not jointly evaluate ECG-conditioned multi-view generation and disease information retention. We introduce ECGEchoDiag, which links multi-view Echo recordings within patients, pairs each recording with synchronized ECG, and provides annotations for eight diseases to evaluate video generation and its patient-level diagnostic value.

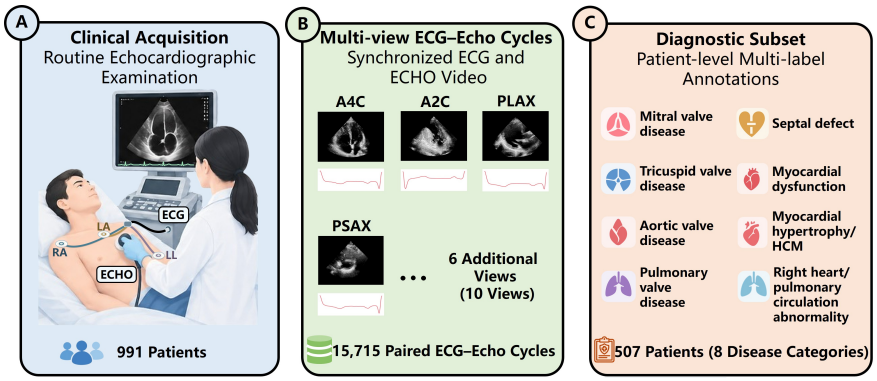


Figure 2: Overview of ECGEchoDiag. (A) ECG and Echo acquisition during routine echocardiographic examinations of 991 patients. (B) 15,715 paired cardiac cycles across 10 views, each comprising an Echo video and a synchronized ECG segment. (C) A diagnostic subset of 507 patients with multi-label annotations for eight cardiac disease categories.

3 ECGEchoDiag Dataset

3.1 Dataset Overview and Cohort Construction

ECGEchoDiag is a dataset for ECG-conditioned multi-view echocardiographic video generation and patient-level disease recognition. As illustrated in Figure 2, it contains 15,715 ECG–Echo paired records from 991 patients, spanning 10 echocardiographic views. Each patient contributes records from a subset of these views, and each record pairs an echocardiographic video with a synchronized ECG signal. The dataset preserves associations across views from the same patient, supporting video generation for individual views and cross-view modeling. Of these patients, 507 have multi-label annotations covering eight cardiac disease categories and form the diagnostic subset. This subset provides disease-related supervision and supports evaluation of disease recognition by jointly analyzing generated videos from multiple views.

3.2 Data Acquisition and Processing

ECGEchoDiag dataset were collected from echocardiographic examinations and synchronized ECG recordings conducted at [Anonymous Hospital] between October 2018 and November 2024, as illustrated in Figure 2(A). All examinations used Philips EPIQ CVx, EPIQ 7C, iE Elite, or iE33 systems. Raw ECHO videos, synchronized ECG signals, and patient information, including dates of birth, were stored in DICOM format. Echocardiographers acquired views such as A4C and PLAX according to each patient's clinical needs, resulting in varying numbers and combinations of views across patients. To protect patient privacy and facilitate model processing, we developed a dedicated program to extract ECG waveforms and ECHO videos from DICOM files, remove information outside the sector-shaped ECHO imaging region, and save both modalities in NPZ format. For each record, two echocardiographers reviewed the ECG signal and Echo video to identify end-diastole and extract one complete cardiac cycle from one end-diastole to the next. Disease annotations were derived from clinical diagnostic records and grouped into the eight cardiac disease categories shown in Figure 2(C), yielding patient-level multi-label annotations. Patients without diagnostic records

Table 2: Dataset statistics. Values with ranges are reported as median [range].

Metric	Full cohort	Diagnostic subset
Patients	991	507
Paired ECG–echo cycles	15,715	8,494
Heart-disease-positive patients	–	481
Heart-disease-negative patients	–	26
Records per patient	12 [1–108]	13 [1–107]
Cardiac-cycle length (frames)	41 [10–435]	39 [10–125]
Echo frame rate (fps)	51.3 [9.2–138.4]	51.3 [11.0–133.1]
ECG sampling rate (Hz)	1,000	1,000
Native echo resolutions (pixels)	640×480, 800×600, 1024×576, 1024×768	

were not assigned negative labels; their paired data remained available for ECHO video generation. The study was approved by the ethics review board of [Anonymous Institution]. All participants provided written informed consent for research use and public release of their de-identified data.

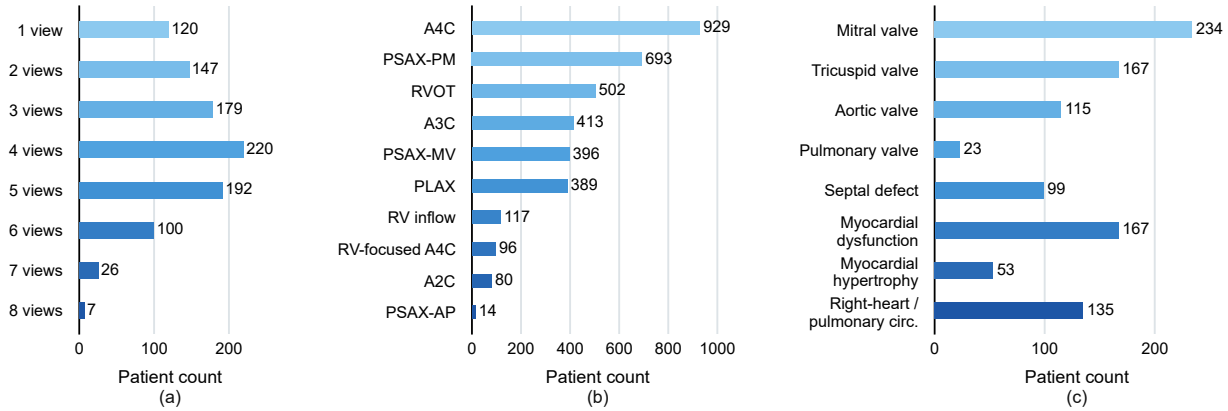


Figure 3: Dataset composition. (a) Views per patient in the cohort. (b) Paired ECG–ECHO cycles across 10 view categories in the cohort. (c) Positive-patient counts for eight non-exclusive disease categories in the diagnostic subset.

3.3 Dataset Statistics

As shown in Table 2, the dataset contains 15,715 paired ECG–ECHO cardiac cycles from 991 patients across 10 supported echocardiographic views. Each sample consists of an ECHO video spanning one complete cardiac cycle and its synchronized ECG segment. Patients have a median of 12 records and 4 distinct views, with the number of available views ranging from 1 to 8. Figure 3(a–b) shows the distribution of available views per patient and the number of patients with each view. The ECHO cycles contain a median of 41 frames at a median frame rate of 51.3 fps. The ECG signals were sampled at 1,000 Hz, and the native videos have four spatial resolutions: 640×480, 800×600, 1024×576, and 1024×768. The diagnostic subset comprises 507 patients with patient-level clinical labels and 8,494 paired cycles. Of these patients, 481 are positive for cardiac disease and 26 are negative. Patients without diagnostic labels are not treated as negative. The eight target disease categories include several types of valvular disease, septal defects, and myocardial abnormalities. Figure 3(c) shows the number of positive patients in each category.

3.4 Data Splitting

To prevent patient-level information leakage, we split the dataset by patient into training, validation, and test sets with a target ratio of approximately $70\% : 15\% : 15\%$. All echocardiographic views, original records, and derived cardiac cycles from each patient remain in the same split. While maintaining the target patient counts, we use a patient-level multilabel balancing strategy that jointly considers positive-patient counts across the eight disease categories, valid-label counts, video sequence counts, and view coverage. The resulting training, validation, and test sets contain 694, 149, and 148 patients, respectively, with 11,005, 2,363, and 2,347 paired ECG–Echo cardiac cycles, corresponding to cycle proportions of $70.03\% : 15.04\% : 14.93\%$. For each disease category, we calculate the proportions of its positive patients assigned to the three splits. Averaging these proportions equally across the eight categories yields

a macro-average allocation of $70.18\% : 15.14\% : 14.68\%$. Comorbidities and limited positive-patient counts in some categories prevent all categories from simultaneously matching the target ratio exactly. The test set is reserved exclusively for final evaluation and is not used for model training, hyperparameter selection, or classification threshold selection.

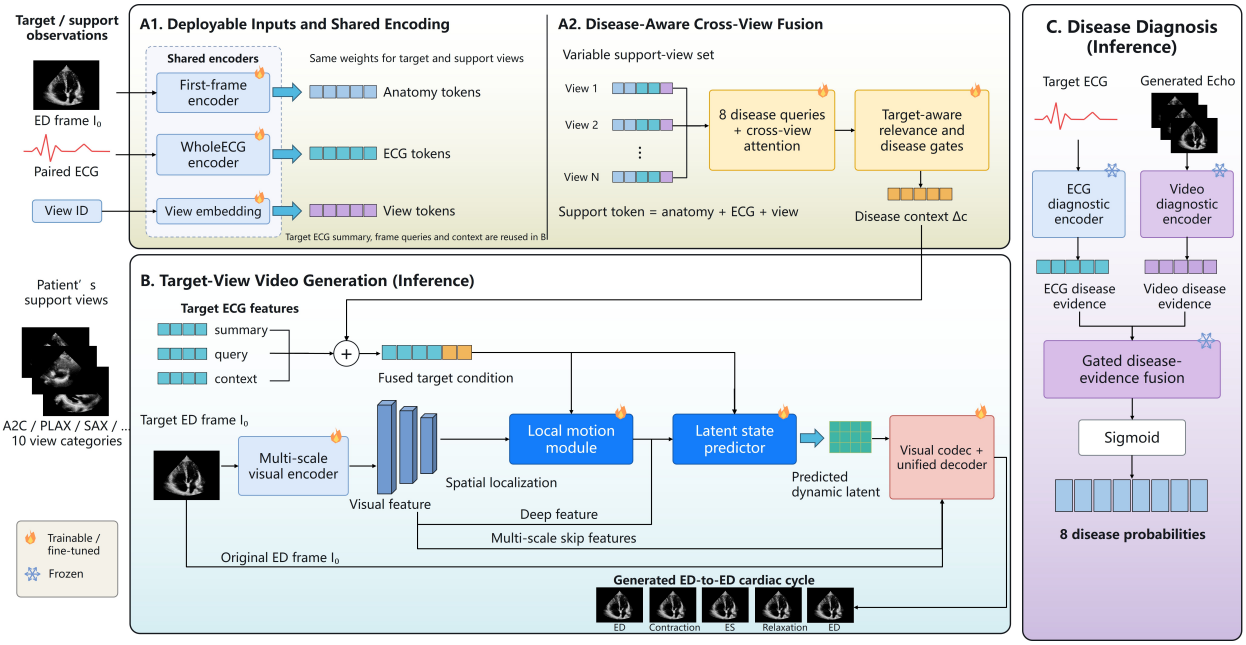


Figure 4: Overview of EchoGenDiag. (A) Shared encoding and disease-aware cross-view fusion extract complementary information. (B) ECG and reference anatomy guide target-view video generation. (C) Gated fusion combines ECG and generated-video evidence for eight-disease recognition. Flames and snowflakes denote trainable and frozen modules, respectively.

4 Method: EchoGenDiag

4.1 Task Formulation and Overview

Generating an Echo video requires reconstructing how cardiac anatomy changes over a heartbeat. A reference frame captures anatomy at one instant, while ECG provides timing cues for inferring cardiac motion. Each view reveals only part of the heart, and other views may provide complementary evidence. EchoGenDiag integrates these observations by selecting relevant support information, using it to guide target-view motion, and reconstructing the video from the reference frame.

Problem Formulation. Given a reference frame I_0^t of target view v_t, its synchronized ECG E^t, and support observations $\mathcal{S}^t = \{(I_0^k, E^k, v_k)\}_{k=1}^K$, we aim to generate a target-view video spanning one cardiac cycle. Each support observation comprises a reference frame, its ECG, and a view identifier. Let $\mathcal{M}$ denote timestamps and the valid imaging sector. We formulate generation as

$$\hat{X}^t = G_\theta(I_0^t, E^t, v_t, \mathcal{S}^t; \mathcal{M}), \qquad \hat{I}_0^t = I_0^t. \tag{1}$$

The reference frame is preserved, and subsequent frames are jointly predicted without access to future real frames from any view. The central challenge is to translate complementary observations into motion appropriate for the target view. Target-aware fusion first selects support evidence conditioned on the target ECG and view identity. To make this evidence guide motion throughout the cycle, a phase-motion adapter combines it with cardiac phase to modulate ECG-conditioned dynamics. Local and global motion modeling translate these cues into spatial changes, while the decoder combines reference-feature warping with content synthesis to reconstruct the video. Since visual similarity alone does not ensure disease-related information preservation, a frozen diagnostic teacher supplements reconstruction objectives with disease-label supervision and prediction consistency between real and generated videos during training. The following subsections detail each component.

4.2 PHASE-ADAPTIVE CROSS-VIEW MOTION CONDITIONING

The aggregated cross-view context is shared across the cardiac cycle and therefore does not explicitly describe how its contribution changes over time. We combine it with the target cycle progress $p_i = (\tau_i - \tau_0)/(\tau_{T-1} - \tau_0)$ to obtain frame-dependent motion modulation. Let $\phi(p_i)$ denote the phase encoding. Re-weighting the target-relevant support representations gives a compact representation d, which is fused with the phase encoding as

$$h_i = F_m([d; F_p(\phi(p_i))]), \qquad r_i = \tanh(g_m)h_i. \tag{2}$$

Here, F_p and F_m denote the phase projection and interaction network. The phase signal does not replace the ECG representation; instead, it organizes how cross-view evidence influences different stages of the target cycle.

The resulting residual is injected into the frame-wise ECG query and temporal context as $\widetilde{Q}_i^t = Q_i^t + \Delta + r_i$ and $\widetilde{C}_i^t = C_i^t + \Delta + \tanh(g_{\text{ctx}})r_i$. After local-state aggregation, a bounded coefficient further modulates the motion-related channels. All newly introduced modulation paths are initialized as zero residuals or identity mappings, preserving the original generation behavior at initialization while allowing the model to gradually learn phase-dependent cross-view corrections.

4.3 FIRST-FRAME-ANCHORED LOCAL MOTION GENERATION

The phase-adapted conditions must be associated with spatial locations in the target view, since different cardiac regions can exhibit distinct motion at the same time. We therefore anchor local motion modeling to the target first frame. A multi-scale image encoder extracts target-view features, and four learnable queries identify soft support regions within the valid imaging sector together with local representations ℓ_j. These regions organize local appearance and motion without being hard-coded to specific chambers or valves. The initial regional state is defined as $H_{j,i}^{(0)} = \ell_j + \widetilde{Q}_i^t + \widetilde{s}^t$. Inter-region attention exchanges information across local regions, while temporal attention updates each region over the cardiac cycle using the real-time bias $B_{i,m} = -|\tau_i - \tau_m|/\ell_+$, where $\ell_+ > 0$ is learnable. The resulting states predict first-frame-relative displacement and feature changes, which are aggregated into a spatially resolved local representation.

To preserve the observed anatomy while allowing new appearance to emerge, the decoder combines first-frame warping with direct content completion. For the i-th predicted frame,

$$\widehat{I}_i^t = (1 - G_i) \odot W(I_0^t, F_i) + G_i \odot D_i, \qquad i = 1, \ldots, T-1, \tag{3}$$

where F_i, D_i, and G_i denote the displacement field, completion image, and pixel-wise fusion gate, respectively. The warping branch propagates patient-specific texture and structure from the observed frame, whereas the completion branch accounts for changes that cannot be represented by geometric transport alone. All local regions jointly contribute to the full-frame reconstruction rather than being generated independently and stitched together.

4.4 JOINT TRAINING OBJECTIVE AND INFERENCE

We jointly optimize generation and diagnostic supervision to constrain visual fidelity, temporal variation, and disease-related information. The generation objective retains the base reconstruction, latent-dynamics, and motion regularization terms, and augments them with a sector-restricted reconstruction loss $\mathcal{L}_{\text{sec}}$ and a temporal-rate loss $\mathcal{L}_{\text{vel}}$:

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{base}} + \lambda_{\text{sec}}\mathcal{L}_{\text{sec}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}. \tag{4}$$

$\mathcal{L}_{\text{sec}}$ applies a Charbonnier penalty within the valid ultrasound sector, while $\mathcal{L}_{\text{vel}}$ compares adjacent-frame intensity changes normalized by their actual time intervals. The observed first frame is excluded from reconstruction loss, whereas its transition to the first predicted frame remains part of the temporal supervision.

To preserve diagnostic information, we employ a frozen teacher T_{diag}. Let $z_g = T_{\text{diag}}(\widehat{X}^t, v_t)$ and $z_r = T_{\text{diag}}(X^t, v_t)$ denote the logits of generated and real videos, respectively. For label vector y and validity mask m, BCE supervision is applied to all valid labels, while teacher consistency is enforced only when the real-video teacher prediction agrees with the available label. We define $\bar{m}_c = m_c \mathbb{I}[(2y_c - 1)z_{r,c} > 0]$, and optimize

$$\mathcal{L}_{\text{diag}} = \omega \left\langle \ell_{\text{BCE}}(z_g, y) \right\rangle_m + (1 - \omega) \left\langle \ell_\beta(z_g - z_r) \right\rangle_{\bar{m}}, \tag{5}$$

where ℓ_β is Smooth-L_1 loss and $\langle\cdot\rangle$ denotes masked averaging. Teacher parameters and real-video outputs are fixed, so gradients are propagated only through the generated-video branch.

After an initial warm-up, the diagnostic term is linearly introduced, giving the final objective $\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{diag}}\eta(s)\mathcal{L}_{\text{diag}}$. At inference, the diagnostic teacher and all training-only references are removed. EchoGenDiag uses only the observed target/support first frames, ECG signals, view identifiers, and auxiliary metadata to generate the target cardiac cycle; no future Echo frames or disease labels are accessed.

Table 3: Main results for ECG-guided echocardiography video generation and diagnosis.

		Reconstruction Fidelity			Diagnostic Fidelity	
Setting	**Method**	FID↓	FVD↓	Sector-SSIM↑	AUPRC↑	AUROC↑
ECG-only	ECG encoder baseline	–	–	–	0.410	0.705
ECG-only	Ours ECG-only	–	–	–	**0.453**	**0.707**
Single-view	First-frame repetition	16.499	1134.358	0.484	0.337	0.659
Multi-view	First-frame repetition	7.001	1022.649	0.482	0.518	0.783
Single-view	ECHOPulse-FT	88.348	1047.469	0.573	0.476	0.756
Multi-view	ECHOPulse-FT	72.111	906.273	0.559	0.455	0.769
Multi-view	ECHOPulse-FT+DS	**71.927**	**901.401**	**0.561**	**0.473**	**0.776**
Single-view	**Ours-SV V21 strict A4C**	49.856	890.256	**0.571**	0.506	**0.788**
Multi-view	**Ours V21 Hybrid-Delta**	**40.722**	**770.646**	0.556	**0.590**	**0.821**
Single-view	*Reference: Real A4C*	0.000	0.000	1.000	0.565	0.788
Multi-view	*Reference: Real multi-view Echo*	0.000	0.000	1.000	0.619	0.830

↓ indicates lower is better, and ↑ indicates higher is better. A dash indicates that the metric is not applicable. DS denotes disease-supervised fine-tuning.

5 BENCHMARK

5.1 EXPERIMENTAL SETUP

Protocol and Implementation. We follow the patient-level split defined in Section 3.4, with no patient overlap across the training, validation, and test sets. Each cardiac cycle is represented by $T = 21$ frames at 128×128 resolution, preserving the end-diastolic reference frame and the cycle endpoint. Frame 0 is provided as the observed condition, while the remaining 20 frames are jointly generated. Model selection, hyperparameter tuning, and disease-classification thresholds are determined exclusively on the validation set, and the test set is reserved for final evaluation. Unless otherwise specified, all models are trained for 8 epochs using AdamW with a learning rate of 5×10^{-5}, weight decay of 0.01, gradient clipping at 1.0, a batch size of 6, and BF16 mixed precision. The diagnostic teacher used during generator training remains frozen throughout optimization. No test labels or future Echo frames are used for model selection or normal generation.

Baselines. We compare against First-frame Repetition, EchoDiffusion-FT, EchoFlow-FT, and ECHOPulse (Li et al., 2025). First-frame Repetition copies the reference frame across all future timestamps. ECHOPulse-ZS uses the released checkpoint without adaptation, while ECHOPulse-FT, EchoDiffusion-FT, and EchoFlow-FT are adapted on our training split. Ours-SV removes the cross-view support pathway, whereas Ours (full) additionally uses available support views from the same examination. For baselines without native cross-view conditioning, multi-view evaluation applies the shared model independently to each view; only Ours (full) explicitly conditions target generation on same-exam support views.

Evaluation Metrics. We evaluate reconstruction fidelity using **FID**, **FVD**, and **Sector-SSIM**, and diagnostic fidelity using macro **AUPRC** and macro **AUROC** from a frozen multi-label classifier trained only on real Echo. Higher Sector-SSIM, AUPRC, and AUROC indicate better performance, whereas lower FID and FVD are preferred.

5.2 OVERALL COMPARISON

Table 3 summarizes the overall performance of ECG-guided Echo generation and disease recognition. Our models achieve strong reconstruction quality in both single-view and multi-view settings, while the full model further improves diagnostic fidelity by exploiting complementary information from same-exam support views. Ours-SV performs strongly among learned single-view generators, whereas the multi-view model maintains comparable reconstruction quality and achieves the best disease-recognition performance among generated-video methods, substantially narrowing the gap to real multi-view Echo. Notably, First-frame Repetition obtains a favorable FID despite containing no cardiac motion, highlighting that frame-level distributional similarity alone may not reflect video-level or diagnostic fidelity. Overall, these results show that combining ECG, reference anatomy, and cross-view information enables EchoGenDiag to better preserve both visual structure and patient-specific diagnostic information.

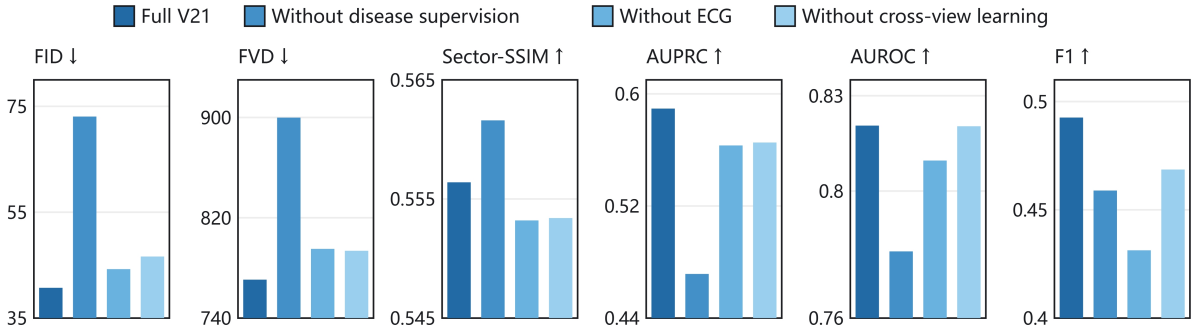


Figure 5: Ablation study of disease supervision, ECG conditioning, and cross-view learning across reconstruction and diagnostic metrics.

5.3 Ablation Studies

All variants use the same patient-level split, input protocol, training schedule, and evaluation setting. Starting from V21 Hybrid-Delta, we individually remove disease supervision, ECG conditioning, and cross-view learning to assess their contributions. As shown in Fig. 5, the complete model provides the best overall balance between reconstruction quality and diagnostic fidelity.

Removing disease supervision causes the largest degradation in both distributional quality and diagnostic performance: FID/FVD increase from 40.72/770.65 to 73.07/899.89, while AUPRC, AUROC, and F1 drop from 0.5895, 0.8206, and 0.4926 to 0.4715, 0.7810, and 0.4589, respectively. Sector-SSIM increases slightly, suggesting that better pixel-level similarity alone does not guarantee preservation of disease-relevant dynamics. Removing ECG conditioning also consistently reduces reconstruction and diagnostic performance, with a particularly clear drop in F1, indicating that ECG provides complementary motion cues beyond the initial Echo frame. Similarly, disabling cross-view learning degrades FID, FVD, Sector-SSIM, AUPRC, and F1, while AUROC remains nearly unchanged. Overall, ECG conditioning and cross-view information provide complementary benefits, whereas disease supervision is critical for retaining clinically discriminative information beyond structural similarity alone.

6 Conclusion

We introduced ECGEchoDiag, a dataset and benchmark for ECG-conditioned multi-view Echo video generation and diagnostic information preservation. It contains 15,715 synchronized ECG–Echo cardiac cycle pairs from 991 patients across 10 views, with eight-disease multi-label annotations for 507 patients. We further proposed EchoGenDiag, combining target-aware cross-view fusion, phase-dependent motion modeling, and reference-guided reconstruction with diagnostic-teacher supervision. Our evaluation examines reconstruction quality, temporal consistency, and patient-level disease recognition, comparing real and generated videos and single-view and multi-view variants. Together, the dataset, baseline, and evaluation protocol support the study of diagnostically informative Echo generation beyond visual fidelity.

AI use statement

Generative AI tools were used solely for language polishing to improve the clarity, grammar, and readability of the manuscript. They were not used to generate thesis content, propose research questions or hypotheses, design the methodology or experiments, implement the proposed methods, process or analyze the dataset, interpret experimental results, or draw scientific conclusions.

All research ideas, methodological designs, experimental procedures, data analyses, results, and conclusions were developed and verified by the authors. All AI-assisted language edits were carefully reviewed by the authors, who take full responsibility for the final content of the manuscript

References

Zachi I Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M McKie, Dorothy J Ladewig, Gaurav Satam, Patricia A Pellikka, Maurice Enriquez-Sarano, Peter A Noseworthy, Thomas M Munger, et al. Screening for cardiac contractile

dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature medicine*, 25(1):70–74, 2019.

Joshua P Barrios, Minhaj U Ansari, Jeffrey E Olgin, Sean Abreau, Jacques Delfrate, Elodie L Langlais, Robert Avram, and Geoffrey H Tison. Multiview deep learning improves detection of major cardiac conditions from echocardiography. *Nature Cardiovascular Research*, 5(3):234–245, 2026.

Donald M Bers. Cardiac excitation–contraction coupling. *Nature*, 415(6868):198–205, 2002.

Chieh-Ju Chao, Jiwoong Jeong, Reza Arsanjani, Kihong Kim, Yi-Lin Tsai, Wen-Chung Yu, Juan M Farina, Ahmed K Mahmoud, Chadi Ayoub, Martha Grogan, et al. Echocardiography-based deep learning model to differentiate constrictive pericarditis and restrictive cardiomyopathy. *Cardiovascular Imaging*, 17(4):349–360, 2024.

Pierre Elias and Joshua Finer. Echonext: a dataset for detecting echocardiogram-confirmed structural heart disease from ecgs. *PhysioNet. doi*, 10, 2025.

Pierre Elias, Timothy J Poterucha, Vijay Rajaram, Luca Matos Moller, Victor Rodriguez, Shreyas Bhave, Rebecca T Hahn, Geoffrey Tison, Sean A Abreau, Joshua Barrios, et al. Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. *Journal of the American College of Cardiology*, 80(6):613–626, 2022.

Illueca Fernandez, T Pollard, BE Moody, L Lehman, B Gow, C Fernandes, C Xie, A Johnson, RG Mark, T Heldt, et al. Heartcycle: A comprehensive dataset of synchronized impedance cardiography and echocardiography for accurate hemodynamic predictions. 2025.

Shinichi Goto, Keitaro Mahara, Lauren Beussink-Nelson, Hidehiko Ikura, Yoshinori Katsumata, Jin Endo, Hanna K Gaggin, Sanjiv J Shah, Yuji Itabashi, Calum A MacRae, et al. Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. *Nature communications*, 12(1): 2726, 2021.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023.

Brian Gow, Tom Pollard, Nathaniel Greenbaum, Benjamin Moody, Ahram Han, Jonathan W Waks, Alistair Johnson, Elizabeth Herbst, Parastou Eslami, Ashish Chaudhari, et al. Mimic-iv-echo: echocardiogram matched subset. Technical report, PhysioNet, 2026.

Gregory Holste, Evangelos K Oikonomou, Márton Tokodi, Attila Kovács, Zhangyang Wang, and Rohan Khera. Complete ai-enabled echocardiography interpretation with multitask deep learning. *Jama*, 334(4):306–318, 2025.

Zhe Huang, Gary Long, Benjamin Wessler, and Michael C Hughes. A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms. In *Machine Learning for Healthcare Conference*, pp. 614–647. PMLR, 2021.

Kenya Kusunose, Takashi Abe, Akihiro Haga, Daiju Fukuda, Hirotsugu Yamada, Masafumi Harada, and Masataka Sata. A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. *Cardiovascular Imaging*, 13(2_Part_1):374–381, 2020.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Shuenn-Yuh Lee, Kai-Ze Lei, Ju-Yi Chen, and Chun-Rong Huang. Descriptor: Cardiac multi-view echo dataset (cameo). *IEEE Data Descriptions*, 3:47–53, 2025.

Yiwei Li, Sekeun Kim, Zihao Wu, Hanqi Jiang, Yi Pan, Pengfei Jin, Sifan Song, Yucheng Shi, Xiaowei Yu, Tianze Yang, et al. Echopulse: Ecg controlled echocardio-gram video generation. In *International Conference on Learning Representations*, volume 2025, pp. 22865–22885, 2025.

Aaron Long, Joshua Finer, Heidi Hartman, Dustin Hartzel, Linyuan Jing, Christopher Kelsey, Daniel Rocha, Jeffrey Ruhl, David Vanmaanen, Youssef Elnabawi, et al. Deep learning for echocardiographic assessment and risk stratification of aortic, mitral, and tricuspid regurgitation: the delineate-regurgitation study. *European Heart Journal*, 46 (28):2780–2791, 2025.

Ahmed M Alaa, Anthony Philippakis, and David Sontag. Etab: A benchmark suite for visual representation learning in echocardiography. *Advances in Neural Information Processing Systems*, 35:19075–19086, 2022.

Masoud Mokhtari, Neda Ahmadi, Teresa SM Tsang, Purang Abolmaesumi, and Renjie Liao. Gemtrans: A general, echocardiography-based, multi-level transformer framework for cardiovascular diagnosis. In *International Workshop on Machine Learning in Medical Imaging*, pp. 1–10. Springer, 2023.

Kabir Hamzah Muhammad, Marawan Elbatel, Yi Qin, and Xiaomeng Li. Echo-path: Pathology-conditioned echo video generation. In *International Workshop on Applications of Medical AI*, pp. 268–277. Springer, 2025.

Emmanuel Oladokun, Sarina Thomas, Jurica Šprem, and Vicente Grau. Echolvfm: One-step video generation via latent flow matching for echocardiogram synthesis. *arXiv preprint arXiv:2603.13967*, 2026.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

Timothy J Poterucha, Linyuan Jing, Ramon Pimentel Ricart, Michael Adjei-Mosi, Joshua Finer, Dustin Hartzel, Christopher Kelsey, Aaron Long, Daniel Rocha, Jeffrey A Ruhl, et al. Detecting structural heart disease from electrocardiograms using ai. *Nature*, 644(8075):221–230, 2025.

Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán Gilligan Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'artagnan: Counterfactual video generation. In *International conference on medical image computing and computer-assisted intervention*, pp. 599–609. Springer, 2022.

Hadrien Reynaud, Mengyun Qiao, Mischa Dombrowski, Thomas Day, Reza Razavi, Alberto Gomez, Paul Leeson, and Bernhard Kainz. Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis. In *International conference on medical image computing and computer-assisted intervention*, pp. 142–152. Springer, 2023.

Hadrien Reynaud, Qingjie Meng, Mischa Dombrowski, Arijit Ghosh, Thomas Day, Alberto Gomez, Paul Leeson, and Bernhard Kainz. Echonet-synthetic: Privacy-preserving video generation for safe medical data sharing. In *International conference on medical image computing and computer-assisted intervention*, pp. 285–295. Springer, 2024.

Hadrien Reynaud, Alberto Gomez, Paul Leeson, Qingjie Meng, and Bernhard Kainz. Echoflow: A foundation model for cardiac ultrasound image and video generation. *arXiv preprint arXiv:2503.22357*, 2025.

Pooneh Roshanitabrizi, Pengfei Guo, Artur Arturi Aharonyan, Kelsey Brown, Taylor Gloria Broudy, Abhijeet Parida, Austin Tapp, Zhifan Jiang, Alison Tompsett, Joselyn Rwebembera, et al. Synthesis of pathological dual-channel color doppler echocardiograms for equitable diagnosis of heart diseases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 589–599. Springer, 2025.

Jonathan S Steinberg, Niraj Varma, Iwona Cygankiewicz, Peter Aziz, Paweł Balsam, Adrian Baranchuk, Daniel J Cantillon, Polychronis Dilaveris, Sergio J Dubner, Nabil El-Sherif, et al. 2017 ishne-hrs expert consensus statement on ambulatory ecg and external cardiac monitoring/telemetry. *Heart rhythm*, 14(7):e55–e96, 2017.

Elias Stenhede, Joanna Sulkowska, Eivind Bjørkan Orstad, Henrik Schirmer, and Arian Ranjbar. Echoxflow: A beamspace echocardiography dataset for cardiac motion, flow, and function. *arXiv preprint arXiv:2605.05447*, 2026.

Alvaro E Ulloa-Cerna, Linyuan Jing, John M Pfeifer, Sushravya Raghunath, Jeffrey A Ruhl, Daniel B Rocha, Joseph B Leader, Noah Zimmerman, Greg Lee, Steven R Steinhubl, et al. rechommend: an ecg-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography. *Circulation*, 146(1):36–47, 2022.

Theo Vos, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet*, 396(10258):1204–1222, 2020.

Milos Vukadinovic, I-Min Chiu, Xiu Tang, Neal Yuan, Tien-Yu Chen, Paul Cheng, Debiao Li, Susan Cheng, Bryan He, and David Ouyang. Comprehensive echocardiogram evaluation with view primed vision language ai. *Nature*, 650(8103):970–977, 2026.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

Jing Wang, Xiaofeng Liu, Fangyun Wang, Lin Zheng, Fengqiao Gao, Hanwen Zhang, Xin Zhang, Wanqing Xie, and Binbin Wang. Automated interpretation of congenital heart disease from multi-view echocardiograms. *Medical image analysis*, 69:101942, 2021.

Junxuan Yu, Rusi Chen, Yongsong Zhou, Yanlin Chen, Yaofei Duan, Yuhao Huang, Han Zhou, Tao Tan, Xin Yang, and Dong Ni. Explainable and controllable motion curve guided cardiac ultrasound video generation. In *International workshop on machine learning in medical imaging*, pp. 232–241. Springer, 2024.

Jiansong Zhang, Xiaying Yang, Xiaoling Luo, and Linlin Shen. Echovdiff: Cardiac-cycle echocardiography video generation from arbitrary single frame. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9040–9050, 2026.

Xinrui Zhou, Yuhao Huang, Wufeng Xue, Haoran Dou, Jun Cheng, Han Zhou, and Dong Ni. Heartbeat: towards controllable echocardiography video synthesis with multimodal conditions-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 361–371. Springer, 2024.

A Appendix

You may include other additional sections here.